Estimating the validity of administrative variables

Bakker, B.F.M.

published in
Statistica Neerlandica. Journal of the Netherlands Society for Statistics and Operations Research
2012

DOI (link to publisher)
10.1111/j.1467-9574.2011.00504.x

document version
Publisher's PDF, also known as Version of record

Statistica Neerlandica (2012) Vol. 66, nr. 1, pp. 8–17
doi:10.1111/j.1467-9574.2011.00504.x

Estimating the validity of administrative variables

Bart F. M. Bakker*

Statistics Netherlands/VU University Amsterdam, P.O. Box 24500, 2490 HA, The Hague, The Netherlands

Administrative data have become more important for both official statistics and academic research. One possible problem with such data is that they are biased and have a low validity. Although this problem is often mentioned in a qualitative respect, the validity is seldom quantitatively measured. This article presents a method to estimate the validity of administrative variables. By applying the classical test theory, the validity can be determined by using linked survey and administrative data which should measure the same concepts. This idea is elaborated with an empirical example in which the construct validity of age, gender, educational attainment and wages is determined simultaneously. A linear structural equations model with a measurement component is used to compute the construct validity. The analyses reveal that educational attainment and wages show some bias, but not higher than the bias found in the survey.

Keywords and Phrases: bias, measurement error, classical test theory, linear structural relationships, register data, data quality.

1 Introduction

The use of administrative data in academic research and official statistics has grown. For example, in recent issues of the Dutch social science journal *Mens en Maatschappij*, over one-third of the articles containing empirical research made use of administrative data (Bakker, 2009). At many statistical bureaus preparations for the 2011 Census are well underway, and more and more countries are making use of administrative data (Valente, 2010). The administrative data, also called administrative registers, are combined by linking and applying micro-integration methods to adjust them and make them more consistent. The outcome of these statistical processes is called a *statistical register* or simply a *register* (Bakker, 2011). In countries where register-based censuses are produced, a growing number of official statistics are based on registers, and quality problems may therefore have a huge impact on the effectiveness of the information infrastructure in society.

*bbkr@cbs.nl/b.f.m.bakker@vu.nl

One problem that may occur when administrative data are used for research or statistics is that the concepts measured do not correspond to the desired concepts. In other words, the validity of the measurement leaves much to be desired. The measurement in administrative data may lack validity for various reasons: the administrative concept may differ substantially from the desired concept; persons, or other entities registered, may have an interest in being registered in a particular way; the administrative register may have a severe administrative delay; the administrative practice of the register keeper may lead to biased entries; or the way the register keeper processes the administrative input may lead to more biased data (Bakker, 2010). The micro-integration process should correct for most errors, but it cannot prevent some errors remaining in the resulting statistical registers.

However, although the problem of validity is often mentioned in qualitative terms, validity is seldom measured in quantitative terms. This article presents a method to estimate how valid register variables are. Based on the classical test theory (see, for instance, Novick, 1966), the assumption is that the measurements of validity can be distinguished from reliability by repeated measurement. Validity can then be determined by using linked survey and register data, which should measure the same concepts, and then repeating the measurement. As it is not always possible to repeat measurements, and it is expensive to do so, it is convenient to conceive the survey and register measurement as two items of the same construct. This idea is elaborated in this article with an empirical example.

The article starts with a short review of the literature on validity and reliability of measurement in registers and surveys. Insight into the concept of validity is enhanced by applying linear structural equation models (Jöreskog and Sörböm, 1996; Kline, 2005) with a measurement component. As an empirical example, the construct validity of age, gender, educational attainment and wages is determined simultaneously by using linked register and survey data. Occupational level is also included in the model, but the validity of this variable cannot be computed because it is only measured in one source. Subsequent sections describe the data of the register and survey used and the results of the data analysis. The last section concludes on the usefulness of the method, discusses the implications for research based on administrative data and suggests future methodological research.

2 Validity and reliability

In the classical test theory (Novick, 1966; Jöreskog and Sörbom, 1996; Kline, 2005), two kinds of measurement errors are distinguished: validity and reliability. According to McCall (2001), reliability refers to whether the measurement procedures assign the same value to a characteristic each time it is measured under essentially the same circumstances. Unreliable measurement leads to random error. To estimate the reliability of a measurement instrument, it is necessary to use it twice (Figure 1). The correlation between the two measures is the estimated reliability: the test–retest reliability. A latent variable is used for the concept to be measured (true score η_1). In

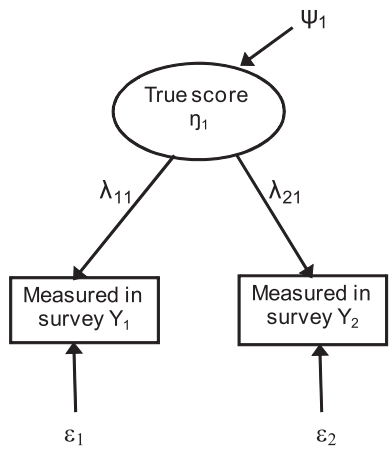

Fig. 1. Estimating the reliability of a survey measure

fact it is measured with $Y_1$and Y_2, variables measured with errors ε_1 and ε_2. The estimated parameters λ_{11} and λ_{21} can be read as factor loadings. Their product equals the test–retest correlation. The higher the λs, the higher the reliability and the lower the error.

Validity refers to how accurately the values assigned in the measurement procedures reflect the actual conceptual variable measured. Invalid measurement leads to systematic error or bias in estimates (McCall, 2001). To estimate the validity of a measurement instrument, the construct validity concept is used (McQueen and Knussen, 2002, 95–98; Singleton and Straits, 2005, 100–105). According to the logic of construct validation, the meaning of a concept is implied by the statements of its theoretical relations to other concepts. The validation process starts with the formulation of the theoretically expected relationships between variables. The more evidence there is in support of the hypothesized relationships, the greater the confidence will be that a particular measurement of the concept is valid.

In this article, a structural equation model is used with a measurement component. Repeated measurement is available for four variables, each taken from both a survey and a register: age, gender, educational attainment and hourly wages. However, in this case the variables are not measured with the same measurement instrument. Therefore, the correlation between the two measurements cannot be read as the test–retest reliability, but as a measure for how differently the measurement instruments measure the concept. Under the condition that the true scores represent the concept well, the factor loadings λ_{11} and λ_{21} can be read as measures of validity and the errors ε_1 and ε_2 as measures of invalidity.

A simple and well-known earnings function model is applied (Figure 2), using the LISREL notation (Jöreskog and Sörbom, 1996). The model is based on, for example, Blau and Duncan (1967), Jencks (1972, 1979), Sewell and Hauser (1980) and for the Netherlands Dronkers and Ultee (1995) and Tolsma and Wolbers (2010). The target population of our study is people with a job of more than 12 hours a

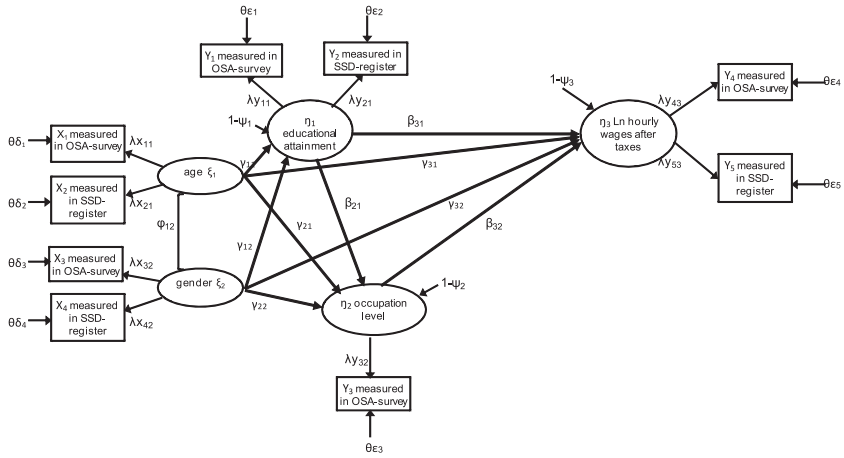

Fig. 2. Model for estimating the validity of register variables age, gender, educational attainment and *ln hourly wages* after taxes

week. Theoretical expectations will have to be formulated in aid of confidence in the outcomes of the model. If available, outcomes of previous research can be used to formulate expectations on the size of the standardized effects. It is expected that age, educational attainment and occupation level have large effects on hourly wages: the expected size is between 0.30 and 0.40. Gender should have a negative effect of approximately −0.20: women earn less than men. Furthermore, the effect of educational attainment on occupation level should be around 0.50, the effects of age and gender on occupation level should be small (approximately 0.10 and −0.10, respectively). The effects of age and gender on education level should be small and positive (both around 0.10).

3 The data

3.1 The survey data

For the survey data, the 'OSA supply panel 2004' (OSA2004) will be used. This is a household sample stratified by age, gender, region and household type. The target population is people aged between 15 and 65 years who are not in daytime education. The survey is panel-based, and respondents from previous waves are approached for a new interview. People who no longer belong to the target population are excluded, and people in sampled households who previously did not belong to the target population are included in the sample if eligible. The interviews took place around 1 October 2004.

Age is measured by asking the date of birth, and calculating the age at interview date. *Gender* is measured by the question: 'What is your gender?'

Educational attainment was measured using the question: 'What is the highest level of education you have completed and for which you have received a certificate?' The respondent could choose between 40 different education programmes on a show card. As these programmes cover different periods, all generations have the option of indicating a suitable education level. This information was harmonized in accordance with the Standard Classification of Education 2006 (SOI).

Occupation level was measured by a rather elaborate questionnaire, asking for job title, main tasks, number of people managed and main managerial tasks. The information was coded into the Netherlands Standard Classification of Occupations 1992 (Bakker, 1993). Occupation level is one of the main criteria of this classification and was derived from the occupational codes.

The wages after taxes are measured by the question 'Can you tell me how much your net wages amount to?'. The interviewer first notes whether the wages are paid weekly, four-weekly, monthly or yearly and then records the amount. The number of working hours is determined through the question 'How many hours do you work according to your employment contract?'. This information always refers to respondent's main job in September 2004. The hourly wages were calculated from the harmonized wages and working hours. This was logarithmically transformed into natural logarithm (ln) *ln hourly wages* after taxes.

3.2 The register data

The register data originate from Statistics Netherlands' Social Statistical Database (SSD) (Bakker, 2002, 2008; Houbiers, 2004). This is a system of linked registers and surveys covering 1999–2010, of which the definite version – adjusted by means of micro-integration – is used. Micro-integration aims at improving quality by harmonizing and completing the data and adjusting them for measurement errors. Micro-integration is executed by applying a set of decision rules, thus transforming administrative register data to statistical register data (Bakker, 2011). This section discusses not only the administrative sources as such, but also some of the micro-integration decision rules used for these four variables. Occupation level is not measured in registers and will not be discussed in this section.

Age and *gender* data are taken from the Population Register. The quality of this information is assumed to be better than that of information from other sources. If people are not included in the Population Register, Statistics Netherlands generally uses information from other sources, but here only people registered in the Population Register are used.

In the Netherlands there is no administrative register for *educational attainment* that covers the entire population, as these administrative registers were only recently first developed. The last time a traditional census that included information about educational attainment was held in the Netherlands was in 1971. These data are not useful for current statistics production, because the respondents can no longer be identified. Therefore, Statistics Netherlands has combined all administrative register

data that are available, for example, the Central Register for Enrolment in Higher Education (available since 1985), the Register of Exam Results including all pupils sitting for final exams in secondary general education from 1999 onwards, the Education Number Registers for secondary general education from 2003 onwards and secondary vocational education from 2005 onwards and a few smaller administrative registers. All these registers are recent and cover only part of the population. People aged 40 years and older, in particular, are not entirely covered.

To complete the population for educational attainment, the Labour Force Surveys (LFS) for 1996–2008 were used. The LFS is a sample survey whose target population is the population aged 15 years and older in the Netherlands, except people living in institutional households. The sample size is just under 1% of the population. School careers are reconstructed using the calendar method. As the LFS is a sample survey, the resulting records are weighted to represent the population not covered by the administrative registers. By combining all information from current administrative registers and surveys, educational attainment can be determined for approximately 45% of the population. In this article, educational attainment measured on 30 September 2004 is used (Bakker, Linder and Van Roon, 2008). By selecting people below 50 years of age, the measured educational attainment is restricted mainly to register entries.

The variable *ln hourly wages* starts with determining the yearly wages before taxes of the main job registered in the fiscal administration. Taxes and insurance contributions are subtracted from the yearly wage before taxes. The wages after taxes of the main job in September 2004 are measured by taking the quotient of the yearly wages after taxes and the number of months the person held the main job. Unfortunately, the working hours of the main job are not registered. Therefore, the working hours from the survey were used to measure hourly wages after taxes, which were logarithmically transformed to get *ln hourly wages*.

3.3 The linked dataset

In the SSD, all registers and surveys are linked to a population backbone. This is a longitudinal version of the Population Register from 1995 onwards. The most important linking variables are a personal identification number (the Dutch Social-Fiscal Number or Citizen Service Number), and the combination of date of birth, gender and address. In some cases surnames are used to link the data. If the data for a person changes, a new entry is made in the population backbone. All records are assigned a linking key if it can be identified in the population backbone. The OSA2004 is linked using name, date of birth, gender and address. The effectiveness is 98.9%: 4730 of the 4782 respondents were assigned a linking key. The records that cannot be linked are not very selective (Fouarge and Grim, 2007); 2873 of the 4730 linked individuals are employees with a job of more than 12 working hours a week.

The register information originates from different administrative registers that differ in linking effectiveness. For most administrative registers, the Social-Fiscal

Number or Citizen Service Number is used to link the data, which leads to an effectiveness of over 97%. Furthermore, many entries are not linked because they do not belong to the population.

The linking key is used to link the OSA2004 to the register data. The effectiveness is almost 100% for gender, age and ln hourly wages. However, for educational attainment the effectiveness is much lower. Moreover, the educational attainment for people aged over 50 years is based mainly on the LFS, and therefore cannot be used to quantify the validity of register information. To restrict the impact of the number of LFS entries, persons younger than 50 years are selected. After these selection processes, only 574 people are eligible to be included in the analysis. To prevent selection bias in the outcomes, the data are weighted by age (in 10-year classes), gender and educational attainment as measured in the OSA2004 survey. If a cell contains fewer than three observations or the weight is over 5.0, it is aggregated with an adjacent cell. The weights are computed with a mean of 1.0.

4 Results

Table 1 shows the correlations between the variables. As expected, variables that are quite obvious measures like age and gender were measured in similar ways in the survey and the register: the correlation is almost 1.00. However, the measurement of educational attainment and ln hourly wages are quite different.

The correlation of educational attainment from survey and register data is 0.768, while the correlation of hourly wage is only 0.823. Moreover, educational attainment measured in the register correlates better with occupation level and hourly wages than the version from surveys. This is true for wages measured in the survey as well as wages measured in the register. However, the differences are small except for the correlations between education and occupation (0.462 for the survey and 0.529 for the register measurement of educational attainment).

Table 1. Correlations between survey and register variables

	Age		Gender		Educational attainment		Occupation level	ln hour wages after taxes	
	Survey	Register	Survey	Register	Survey	Register	Survey	Survey	Register
Age from survey	1.000								
Age from register	0.998	1.000							
Gender from survey	−0.070	−0.072	1.000						
Gender from register	−0.071	−0.073	0.999	1.000					
Educational attainment from survey	−0.133	−0.135	0.037	0.038	1.000				
Educational attainment from register	−0.219	−0.218	0.009	0.010	0.768	1.000			
Occupational level from survey	0.005	0.004	−0.092	−0.091	0.462	0.529	1.000		
ln hourly wages after taxes from survey	0.210	0.211	−0.216	−0.217	0.406	0.427	0.514	1.000	
ln hourly wages after taxes from register	0.314	0.313	−0.188	−0.188	0.298	0.313	0.447	0.823	1.000

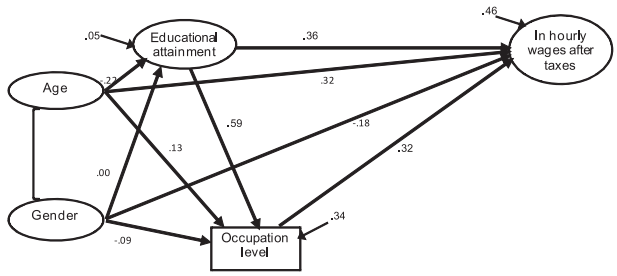

Fig. 3. Evaluating the plausibility of the parameters in the model
Note: The explained variance is shown for each endogenous variable $(1-\phi)$.

The validity of register variables should be demonstrated by applying a structural equation model. The complete estimated model is shown in Figure 2. This model fits the data with a χ^2 of 48 by 18 degrees of freedom. There are no important residual correlations: there is only a small residual correlation between both age variables and both wage variables and one between age and education from registers. The fit of the model did not improve much by adding parameters. Therefore, the model was accepted.

In the next step the plausibility of the estimated parameters in the model is evaluated (Figure 3). If the values of the parameters are implausible, then nothing could be concluded about the validity of the measured variables. However, the results corresponded with our expectations. Educational attainment (0.36), occupation level (0.32) and age (0.32) have large positive effects on wages, while gender has a moderate negative effect (−0.18). Occupation level is affected by educational attainment (0.59). Age has a moderate negative effect on educational attainment (−0.22). Selecting people younger than 50 years could have an effect on the size of this negative effect, if the relationship between age and educational attainment is not linear. However, inspection of our data does show a linear relationship. All other parameters are small as expected.

Lastly, the measurement errors are evaluated (Table 2). The relationship between ε and λ is $\varepsilon = 1-\lambda^2$. The higher the λ and the lower the ε, the higher the validity of the measurement. $\lambda = 1$ indicates perfect measurement without error. The measurement errors of age and gender are very small and not significant, so they are measured almost perfectly in both the survey and the register. However, the errors in educational attainment and ln hourly wages are large and significant. Educational attainment is measured with less error in the register than in the survey. The survey

Table 2. Measurement errors in survey and register variables

	OSA survey			SSD register		
	λ	ε	Significance	λ	ε	Significance
Age	1.00	0.00		1.00	0.00	
Gender	1.00	0.00		1.00	0.01	
Educational attainment	0.82	0.33	**	0.95	0.11	**
Occupational level	–	–		–	–	
ln hourly wages after taxes	0.95	0.10	**	0.87	0.24	**

Note: Significant $p<0.01$

measurement has a significant error of 0.33, while the register measurement has an error of only 0.11. For ln hourly wages, the survey is the better measurement. While the register measurement has an error of 0.24, the measurement error for the survey variable is 0.10. The differences between the size of the measurement errors is significant in both cases.

5 Conclusion and discussion

Despite the increased use of administrative register data in academic research, not a lot is known about the quality of these data. This article describes a method to estimate the validity of register data. With the aid of the classical test theory and linear structural equations models, it is possible to quantify the construct validity. Measures from surveys can be linked to measures from registers, and under the condition that the model produces plausible results, the measurement errors can be read as a measure for the validity.

This model was applied to an earnings function, in which age, gender, education level and *ln hourly wages* were measured in a survey and in a register. Occupation level is also part of the model, but it is only measured in the survey. As the model produces plausible results, the estimated measurement errors can be used as a measure for validity. The measurement of educational attainment was better in the register than in the survey. For *ln hourly wages*, the opposite is true.

This article shows that the proposed method is usable for qualitative research of register data. Of course, one of its weaknesses is that the results depend on the knowledge of the relationships of the measured variables and other concepts. In this case there is a thorough theoretical and empirical knowledge of these relationships, grounded in different disciplines like economics and sociology. However, in cases where there is less knowledge, it will be more difficult to apply the method. In general, it would be more difficult to apply it in a new field of research in which concepts and measurement still have to be developed.

Furthermore, it is too early for general conclusions about the quality of register data. This is the first attempt to estimate the validity of some register data. The method will have to be applied to more register data to arrive at more general conclusions. The resulting picture is expected to be mixed: some variables are better measured in a particular register, others in a particular survey. It should also be mentioned that the SSB register has been created by means of micro-integration. Some measurement errors had been detected and adjusted before the data were used in this study. In general, the validity of the register data could be expected to be lower if the original data had been used. The method presented in this article could also be used to determine the validity of the original administrative data. The knowledge of this quality aspect of the original data could be used for the design of the decision rules and therefore to improve the micro-integration process. In the end, it is still worthwhile to apply this method to the improved dataset to estimate the validity of its variables, because it is a quality indicator.

The measurement of educational attainment in the register is hybrid: most of the entries come from administrative registers, but it is completed with entries from sample surveys. This also demonstrates the inconvenience of some register data: sometimes a variable is entirely or partly missing. This urges the researcher to use survey measures to estimate the desired relationships. Moreover, because the character of this variable is hybrid, its estimated validity could not be used for a conclusion on the quality of register data alone.

References

Bakker, B. F. M. (1993), The development of the Standard Classification of Occupations 1992, *Netherlands Official Statistics* **8**, 5–22.

Bakker, B. F. M. (2002), Statistics Netherlands' approach to social statistics: the social statistical dataset, *OECD Statistics Newsletter* **2**, 4–6.

Bakker, B. F. M. (2008), De stand van het Sociaal Statistisch Bestand, *Bevolkingstrends* **56**, 14–18.

Bakker, B. F. M. (2009), *Trek alle registers open!*, Vrije Universiteit, Amsterdam.

Bakker, B. F. M. (2011), *Micro-integration*, Statistics Netherlands, The Hague/Heerlen.

Bakker, B. F. M., F. Linder and D. van Roon (2008), *Could that be true? Methodological issues when deriving educational attainment from administrative datasources and surveys*. Paper prepared for the IAOS Conference on Reshaping Official Statistics Shanghai, 14–16 October.

Blau, P. M., and O. D. Duncan (1967), *The American occupational structure*, Wiley, New York/London/Sydney.

Dronkers, J. and W. C. Ultee (eds) (1995), *Verschuivende ongelijkheid in Nederland*, Van Gorcum, Assen.

Fouarge, D. and R. Grim (2007), *Koppeling van het OSA-Arbeidsaanbodpanel aan administratieve gegevens: verslag en documentatie*, OSA, Tilburg.

Houbiers, M. (2004), Towards a Social Statistical Database and unified estimates at Statistics Netherlands, *Journal of Official Statistics* **20**, 55–75.

Jencks, C. (1972), *Inequality: a reassessment of the effect of family and schooling in America*, Basic Books, New York.

Jencks, C. (1979), *Who gets ahead?* Basic Books, New York.

Jöreskog, K. and D. Sörbom (1996), *LISREL: 8. User's reference guide*, Scientific Software International, Chicago, IL.

Kline, R. B. (2005), *Psychological testing: a practical approach to design and evaluation*, Sage, New York.

McCall, R. B. (2001), *Fundamental statistics for behavioural sciences*, Wadsworth, Belmont.

McQueen, R. and C. Knussen (2002), *Research methods for social science. An introduction*, Prentice Hall, Harlow.

Novick, M. R. (1966), The axioms and principal results of classical test theory, *Journal of Mathematical Psychology* **3**, 1–18.

Sewell, W. H. and R. M. Hauser (1980), The Wisconsin Longitudinal Study of social and psychological factors in aspirations and achievement, *Research in Sociology of Education and Socialization* **1**, 59–99.

Singleton, R. A., JR. and B. C. Straits (2005), *Approaches to social research*, Oxford University Press, Oxford/New York.

Tolsma, J. and M. H. J. Wolbers (2010), *Naar een open samenleving? Recente ontwikkelingen in sociale stijging en daling in Nederland*, RMO, Den Haag.

Valente, P. (2010), *Main results of the UNECE/UNSD Survey on the 2010/2011 round of censuses in the UNECE region*, Eurostat, Luxembourg.

Received: 1 June 2011 Revised: 26 September 2011